BITS F464 Machine Learning

Assignment 2

Date : 9th April, 2020

Team Members

Durba Satpathi	2019A7PS0972H
R Adarsh	2019A7PS0230H
Navdeep Singh	2017B5A71675H

Problem 2A Logistic Regression

1.1. Model Description and implementation

Dataset given:

The given dataset consists of 1372 instances and each input instance has 4 attributes with output being 0 or 1.

Data Processing:

1) The data has been split into 70% (train set) and 30% (test set)
2) 10 independent such random splits have been created and the model is trained each time and the accuracies and loss for each split is obtained and finally the average accuracy and average loss is reported.

Loss Function:

The loss function used in case of logistic regression is,

Loss(h(x),y) = -y log (h(x)) - (1-y)log(h(x))

So, the cost function is summation of loss for all training examples, i.e

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)})$$

$$J(\theta) = \frac{1}{m}[\sum_{i=1}^{m} -y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))]$$

where m is the number of training samples.

Optimization:

Our objective is to find the optimum set of parameters that minimize the cost function. We have Batch Gradient Descent and Stochastic Batch gradient descent as optimization algorithms.

A) Standard/ Batch Gradient Descent:

1. We start by initializing random weights
2. In each iteration we update the parameters,

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

$$\theta_j \leftarrow \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

3. By iterating over the training examples until convergence we are able to reach the optimum parameters that minimize the cost.

B) Stochastic Gradient Descent:

Usually when we deal with bigger datasets, Gradient Descent turns out to be slow computationally since for every iteration it passes the whole dataset to calculate the gradient of loss. Hence we use SGD in which we compute the gradient for just a single example at each iteration. This brings down the computation time.

1. Pick a single data point
2. Compute gradient over that single point
3. Update parameters

1.2. Most important feature in the dataset

The k dimensional weight vector can be used to get **feature importance**. Large positive values of wj signify higher **importance** of the jth **feature** in the prediction of positive class. Large negative values signify higher **importance** in the prediction of negative class.
Based on the values of w (from both gd and sgd) we conclude that attribute 1 is the most important feature because it has the highest value.

1.3. The final train and test metrics

GD

```
0.001
Split 1:
w [-4.03814594 -2.36837703 -2.84873404 -0.37585526]
Split 2:
w [-4.01134896 -2.25313326 -2.74407284 -0.29820204]
Split 3:
w [-3.91876533 -2.26351902 -2.74909806 -0.23488672]
Split 4:
w [-3.93571441 -2.26296208 -2.75177194 -0.23801738]
Split 5:
w [-4.106325   -2.28762003 -2.78688461 -0.32822466]
Split 6:
w [-4.1907308  -2.25397539 -2.82576504 -0.21911847]
Split 7:
w [-4.05642904 -2.23137048 -2.7372538  -0.24187552]
Split 8:
w [-4.23089779 -2.25888622 -2.78655449 -0.06836492]
Split 9:
w [-4.16975544 -2.23819274 -2.81858121 -0.18555894]
Split 10:
w [-4.04136625 -2.16547308 -2.70007749 -0.18188605]
Overall Training Accuracy = 0.9886458333333333
Overall Test Accuracy = 0.9895631067961166
Overall Training f = 0.9871628145287117
Overall Test f = 0.9885396784076953
Overall Training p = 0.9905631398105884
Overall Test p = 0.9909951585879841
Overall Training r = 0.9837894378774885
Overall Test r = 0.9861393861348813
0.005
Split 1:
```

```
w [-7.41369968 -3.95291112 -4.94714727 -0.42641126]
Split 2:
w [-8.14375942 -4.3896455  -5.28244295 -0.60433181]
Split 3:
w [-7.78130657 -4.1767159  -5.17428149 -0.56789485]
Split 4:
w [-7.13337414 -3.89028639 -4.98900349 -0.58280611]
Split 5:
w [-7.52235594 -4.02943026 -5.16324553 -0.64652441]
Split 6:
w [-7.92582615 -4.08887682 -5.35520716 -0.71211684]
Split 7:
w [-7.54332428 -4.04776216 -4.94194046 -0.66210248]
Split 8:
w [-8.71285737 -4.46969575 -5.84985135 -0.58228055]
Split 9:
w [-7.30288293 -3.93597903 -5.01130422 -0.56255782]
Split 10:
w [-7.81261741 -4.08714485 -5.11267025 -0.49789964]
Overall Training Accuracy = 0.9903124999999999
Overall Test Accuracy = 0.9902912621359222
Overall Training f = 0.9891274073594968
Overall Test f = 0.9889950804244126
Overall Training p = 0.9890190743614056
Overall Test p = 0.9884927700621567
Overall Training r = 0.9892535419133701
Overall Test r = 0.989538655314937
0.01
Split 1:
w [-13.51617518  -7.3037016   -9.13045266  -1.16914464]
Split 2:
w [-14.39400949  -7.45518743  -9.43999823  -0.56732084]
Split 3:
w [-13.74840318  -7.36598297  -9.1934849   -1.28755625]
Split 4:
w [-14.13720655  -7.71651412  -9.56671058  -1.25108306]
Split 5:
w [-13.12519795  -7.01503489  -8.78397081  -1.44613712]
Split 6:
w [-15.86278871  -7.99632879 -10.64802135  -1.24560781]
Split 7:
w [-12.4295177   -6.7718341   -8.44841568  -1.24065164]
Split 8:
w [-14.33340001  -7.24530516  -9.5573366   -1.09187058]
Split 9:
w [-15.5655499   -8.63663453 -10.69416156  -1.49676717]
Split 10:
w [-13.03023147  -7.59478871  -8.24980828  -1.15956828]
Overall Training Accuracy = 0.9885416666666667
Overall Test Accuracy = 0.9885922330097087
Overall Training f = 0.9870416647543268
Overall Test f = 0.9871995788477733
Overall Training p = 0.9872645475387731
Overall Test p = 0.9908345416300616
Overall Training r = 0.9868258846106646
```

Overall Test r = 0.9837043560009923

SGD

0.001

Split 1:
w [-2.07322215 -1.25968054 -1.44311017 -0.19464265]
Split 2:
w [-2.08613714 -1.27942906 -1.45703299 -0.22051318]
Split 3:
w [-2.16424168 -1.23775587 -1.42378575 -0.18851362]
Split 4:
w [-2.12489683 -1.24493801 -1.44188796 -0.18940908]
Split 5:
w [-2.09612842 -1.24771077 -1.42023732 -0.18718278]
Split 6:
w [-2.18627763 -1.24060298 -1.43850486 -0.16191276]
Split 7:
w [-2.14089873 -1.26190733 -1.44586332 -0.1898934]
Split 8:
w [-2.05466075 -1.2888392 -1.44897074 -0.29623509]
Split 9:
w [-2.12356713 -1.21984802 -1.42722533 -0.16531652]
Split 10:
w [-2.14071537 -1.24031214 -1.47365855 -0.12934363]
Overall Training Accuracy = 0.9885416666666668
Overall Test Accuracy = 0.9883495145631068
Overall Training f = 0.9872539948531862
Overall Test f = 0.9867537995990538
Overall Training p = 0.9902246654198216
Overall Test p = 0.9910559395459793
Overall Training r = 0.9843062477679224
Overall Test r = 0.9825303453916799

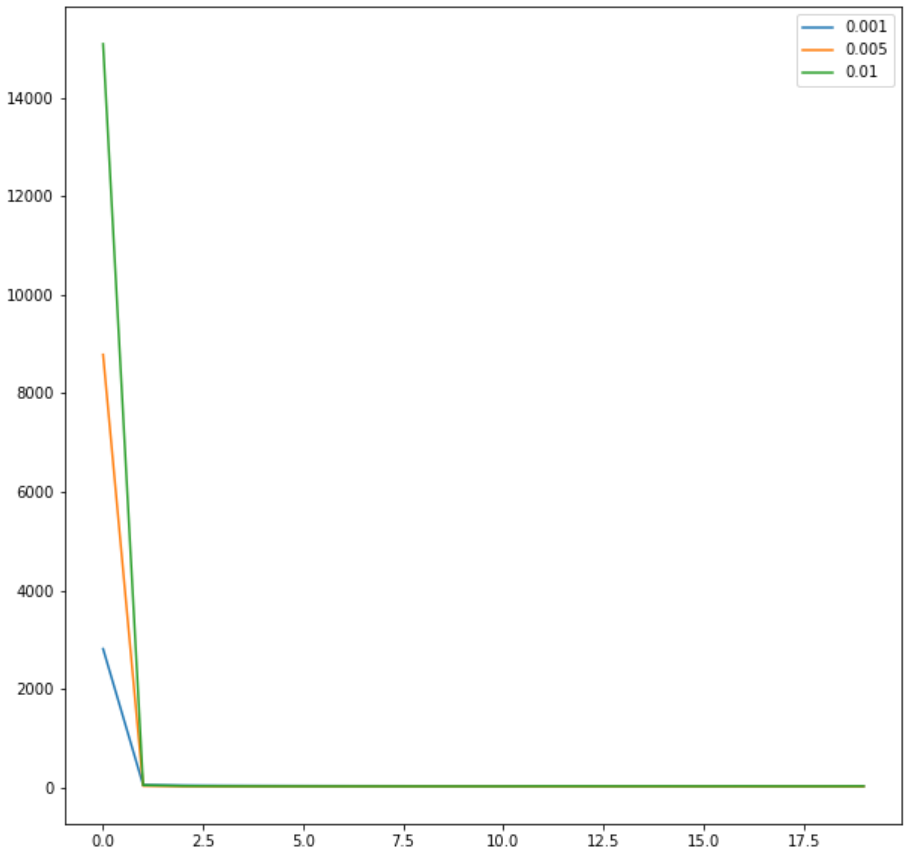
0.005

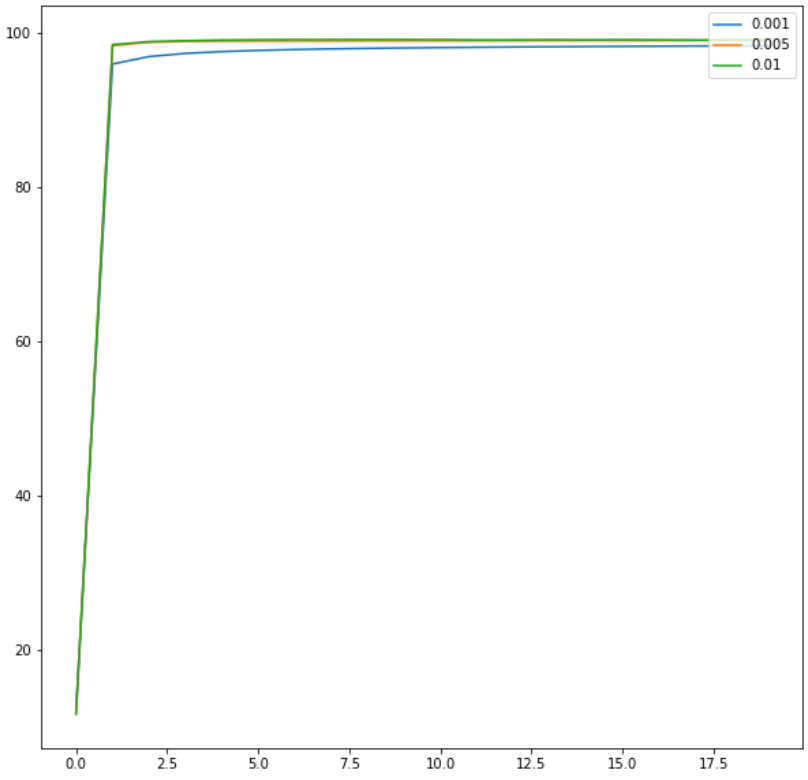
Split 1:
w [-3.54053837 -1.81635802 -2.2898279 -0.12624348]
Split 2:
w [-3.28234026 -1.85982741 -2.29337145 -0.22713803]
Split 3:
w [-3.38296869 -1.93623801 -2.37100814 -0.19716667]
Split 4:
w [-3.1774614 -1.817847   -2.23926619 -0.152745]
Split 5:
w [-3.16816328 -1.85393331 -2.22587139 -0.19568237]
Split 6:
w [-3.41369725 -1.84600753 -2.2567534 -0.14635354]
Split 7:
w [-3.21367825 -1.7330662 -2.13470046 -0.17341631]
Split 8:
w [-3.1252957 -2.03290015 -2.32474377 -0.3060242]
Split 9:
w [-3.40120566 -1.91806741 -2.25895636 -0.26931787]
Split 10:
w [-3.403082   -1.87298065 -2.31049652 -0.18732615]
Overall Training Accuracy = 0.9893750000000001

Overall Test Accuracy = 0.9912621359223301
Overall Training f = 0.9881240655542071
Overall Test f = 0.9901028115530197
Overall Training p = 0.992274365077351
Overall Test p = 0.9944627841643859
Overall Training r = 0.9840152920311878
Overall Test r = 0.9858884358708229
0.01
Split 1:
w [-3.93264128 -2.15803339 -2.68587101 -0.23691526]
Split 2:
w [-3.78894003 -2.07168948 -2.64738459 -0.08579376]
Split 3:
w [-4.02440887 -2.1824088 -2.7340975 -0.17559082]
Split 4:
w [-4.05148133 -2.23301054 -2.80361117 -0.31148475]
Split 5:
w [-4.03233114 -2.20466273 -2.79492661 -0.24570027]
Split 6:
w [-3.97759527 -2.09188383 -2.59183689 -0.11560486]
Split 7:
w [-4.25892209 -2.33779841 -2.99097048 -0.14330594]
Split 8:
w [-4.12299782 -2.23884487 -2.7113735 -0.29344661]
Split 9:
w [-4.29001543 -2.25070721 -2.75868219 -0.02574624]
Split 10:
w [-4.2055247 -2.35879818 -2.90685821 -0.22339015]
Overall Training Accuracy = 0.9892708333333333
Overall Test Accuracy = 0.9890776699029127
Overall Training f = 0.988093993756932
Overall Test f = 0.9875462815339251
Overall Training p = 0.9921160818328948
Overall Test p = 0.9895332284467235
Overall Training r = 0.9841156847233197
Overall Test r = 0.9856507198908135

1.4. Plots of accuracy for three different learning rates using GD and SGD

Gradient Descent





Stochastic Gradient Descent